Information Fusion using the Kalman Filter based on Karhunen-Loève Decomposition

Zhiming Lu[1], Dongxiao Zhang[2], and Yan Chen[3]

[1] Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87645, USA, zhiming@lanl.gov
[2] Department of Civil and Environmental Engineering, and Mork Family Department of Chemical Engineering and Material Sciences, University of Southern California, Los Angeles, CA 90089, USA, donzhang@usc.edu
[3] Mewbourne School of Petroleum and Geological Engineering, University of Oklahoma, Norman, OK 73019, USA, yan@ou.edu

1 Introduction

Properties of porous media, such as hydraulic conductivity and porosity, are intrinsically deterministic. However, due to the high cost associated with direct measurements, these properties are usually measured only at a limited number of locations. The number of direct measurements is definitely not enough to infer the true parameter fields. This problem is further complicated by the fact that medium properties exhibit a high degree of spatial heterogeneity. This combination of significant spatial heterogeneity and a relatively small number of direct observations leads to uncertainty in characterizing medium properties, which in turn results in uncertainty in estimating or predicting the corresponding system responses (such as hydraulic head). Fortunately, it is relatively easy to measure the system responses, which can be used to infer medium properties. With newly developed measurement techniques such as remote sensing and *in-situ* permanent sensors, more observations on system responses become available.

Large efforts have been made to take the advantage of all available observations, both the limited number of direct measurements of medium properties and a larger amount of observations of system responses, to obtain better estimates of the primary parameters of the porous media, thereby reducing the uncertainty associated with model predictions. In hydrology, many inverse models have been developed for aquifer characterization [29, 19, 4, 2, 14, 31, 37, 12, 18, 35, 23]. Review and comparison of some of these inverse models can be found in [36, 11, 26, 41]. In these models, the best estimate of a medium property is obtained by minimizing the mismatch between the estimated values and observed ones. However, these models are not capable of incorporating measurements dynamically, which means that,

if new data are available at a time, one has to run the inverse models again from the very beginning.

Recently, a number of sequential inverse methods (history matching in petroleum enginnering) have been proposed, which is basically extended from the optimal control theory of Kalman filter. These methods include Extended Kalman filter (EKF), Ensemble Kalman filter (EnKF), and their variants. The common ground of these methods is their capability of incorporating observations sequentially, which results in a significant reduction of the number of data dealt at any given time. An up-to-date system, consisting of a best estimate and the corresponding uncertainty represented by a state error covariance matrix, is always provided by these sequential methods. The fundamental difference among these methods is the way by which the state error covariance matrix is estimated. The EKF uses the first-order linearization approach and needs to keep track of the whole covariance function, which is computationally expensive for large-scale problems. The EnKF employs the Monte Carlo method, in which the covariance matrix is updated from a small-sized ensemble (a small number of realizations). Some literature shows that for large-scale highly nonlinear problems the EnKF is superior to the EKF in terms of both efficiency and accuracy [6, 17, 7].

The EnKF is conceptually simple, easy to implement, and capable of accounting for different types of models and observation noises, and its computational cost is relatively low compared to other approaches. The method has been used in a large number of applications in various fields such as meteorology, oceanography, hydrology, and reservoir engineering [6, 7, 13, 20, 28, 27]. The size of the ensemble is crucial for the EnKF, because the standard deviation of the sampling errors of the Monte Carlo method converges very slowly at a rate inversely proportional to the square root of the sample size. To reduce the sampling error by a factor of two, the number of Monte Carlo simulations has to be increased by a factor of four. If the observations are not perfect, the sampling error also appears in the ensemble of observations, since the EnKF needs a set of perturbed observations to model the observation noise. In general, a small ensemble induces a large sampling error, which tends to underestimate the state error covariance; a large ensemble leads to computational inefficiency.

Many methods have been proposed to reduce the sampling error associated with the small-sized ensemble in the EnKF. Ensemble square root filter attempts to avoid the perturbed observations by using different Kalman gains to update the ensemble mean and the deviation from the ensemble mean [34]. Anderson and Anderson [1] used a parameter to broaden the ensemble spread. Double Ensemble Kalman filter divides the ensemble into two parts, and uses the Kalman gain calculated from one ensemble to update the other. These methods showed promising results even with relatively small number of ensemble members. On the other hand, these methods are application-dependent and some parameters are difficult to quantify. Extra effort may need to tune these parameters in order to obtain reasonable results.

For high-resolution, large-scale problems, the dimension of the state in general is fairly large. One major problem of various Kalman filter methods comes down to how to efficiently approximate the state error covariance function in each update with a dimension-reduced approach. The EnKF is one type of dimension-reduced approaches since the number of ensemble members is in general smaller than the dimension of the state. There are other attempts that estimate the state error covariance matrix by selecting some principal modes of the state vector, for instance, singular evolutive extended Kalman filter (SEEK) [30], reduced rank square root filter (RRSQRT) [33], and partially orthogonal ensemble Kalman filter (POEnKF) [16].

Recently, Zhang et al. [40] developed a KL-based Kalman filtering scheme (KLKF), which is a type of dimension-reduced Kalman filtering method based on the Karhunen-Loève (KL) expansion of a medium property and orthogonal polynomial decompositions of dependent variables. In the KLKF method, the covariance of the medium property is efficiently approximated by a small set of eigenvalues and eigenfunctions attributed to the mean square convergence of the KL decomposition. Compared to the full covariance matrix, the finite number of modes used to approximate the covariance represents a significant reduction in random dimensions. In this updating procedure, the forward problem is solved with a moment method based on Karhunen-Loève decomposition (KLME), from which the mean and covariance of the state variables can be constructed, when needed. The statistics of both the medium properties and system responses are then updated with the available measurements at the current time using the auto- and cross-covariance obtained from the forward step. They used a synthetic 2-D example to demonstrate the capability of this new method for a stationary conductivity field and compared the results with those from the EnKF method. Their numerical results show that the KLKF method is capable of significantly reducing the required computational resources with satisfactory accuracy, which indicates the potential applicability of this approach to high-resolution, large-scale predictive models.

This chapter is organized as follows: The problem being addressed is described in detail in Section 2. Section 3 provides some basic concepts and formulations of the data assimilation methods. The Karhunen-Loève decomposition for both stationary and nonstationary hydraulic conductivity fields is given in Section 4. The KL-based moment method for solving the first-oder head and head covariance is outlined in Section 5. In Section 6, the KL-based Kalman filtering scheme is introduced, including a detailed description on updating hydraulic conductivity, head field, and eigenvalues of eigenfunctions of the conditioning covariance. The KLKF method is illustrated using a synthetic example in Section 7, with a detailed discussion on the accuracy and efficiency of the method as compared with the EnKF method. The chapter concludes in Section 8 with a short summary and discussion.

2 Statement of the Problem

We consider transient fluid flow in a heterogeneous porous medium satisfying the following governing equation:

$$\nabla \cdot [K_s(\mathbf{x})\nabla h(\mathbf{x},t)] + g(\mathbf{x},t) = S_s \frac{\partial h(\mathbf{x},t)}{\partial t}, \tag{1}$$

subject to initial and boundary conditions:

$$h(\mathbf{x},0) = H_0(\mathbf{x}), \qquad \mathbf{x} \in D, \tag{2}$$

$$h(\mathbf{x},t) = H(\mathbf{x},t), \qquad \mathbf{x} \in \Gamma_D, \tag{3}$$

$$K_s(\mathbf{x})\nabla h(\mathbf{x},t) \cdot \mathbf{n}(\mathbf{x}) = -Q(\mathbf{x},t), \qquad \mathbf{x} \in \Gamma_N, \tag{4}$$

where $K_s(\mathbf{x})$ is the hydraulic conductivity, $h(\mathbf{x},t)$ is the pressure head, $g(\mathbf{x},t)$ is the source/sink term, S_s is the specific storage, $H_0(\mathbf{x})$ is the initial head in the domain D, $H(\mathbf{x},t)$ is the prescribed head on Dirichlet boundary segments Γ_D, $Q(\mathbf{x},t)$ is the prescribed flux across Neumann boundary segments Γ_N, and $\mathbf{n}$ is an outward vector normal to the boundary $\Gamma = \Gamma_D \cup \Gamma_N$. Here the hydraulic conductivity is considered as a spatially-correlated random variable, while specific storage S_s is treated as a deterministic constant because of its relatively small variability. For simplicity, it is assumed that both initial and boundary conditions are deterministic.

Since K_s is a random function, the flow equations become stochastic partial differential equations, which can be solved through several approaches, such as Monte Carlo simulations, the moment-equation approach [38], and the KL-based moment method (KLME) [39, 21, 22]. All Kalman filter methods need to solve the forward problem, i.e., (1)-(4). The EnKF uses the Monte Carlo simulations to obtain the covariance of the state variable (pressure head), while the KLKF uses the KLME to conduct forward modeling. It has been shown that the KLME method is computationally more efficient than the Monte Carlo simulations [22], which makes it possible to develop an efficient algorithm that takes advantages of both the Kalman filter and the KLME methods.

The hydraulic conductivity is assumed to follow a log normal distribution (see a detailed review in [32]), and we work with the log-transformed variable $Y(\mathbf{x})$, given as:

$$Y(\mathbf{x}) = \ln K_s(\mathbf{x}) = \langle Y(\mathbf{x}) \rangle + Y'(\mathbf{x}), \tag{5}$$

where $\langle Y(\mathbf{x}) \rangle$ is the ensemble mean of $Y(\mathbf{x})$, representing a relatively smooth unbiased estimate of the unknown random function $Y(\mathbf{x})$, and $Y'(\mathbf{x})$ is the zero-mean fluctuation.

Suppose we have N_Y direct measurements of the log hydraulic conductivity $Y_1, Y_2, \cdots, Y_{N_Y}$, taken at locations $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_Y}$, and N_h pressure head measurements located at $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \cdots, \boldsymbol{\chi}_{N_h}$, measured at some time intervals. As mentioned earlier, many inverse models are not capable of dynamically

updating parameter fields. In the following sections, we describe an efficient algorithm to characterize medium properties by dynamically incorporating these measurements when they become available. During this data assimilation process, the stochastic differential flow equation is solved forward with time; while the mean and fluctuation of the hydraulic conductivity are modified together with the computed pressure head to honor all the observations at various times.

3 Basic Concepts of Data Assimilation

Data assimilation is a dynamically-updating process, in which estimates of model parameters and state variables are updated by requiring consistency with observations and governing flow equations. With newly developed measurement techniques such as remote sensing and *in-situ* permanent sensors, more observations become available. However, observations from these techniques are usually indirectly related to model parameters, which are of the most interest in hydrological applications. Moreover, measurements may be corrupted by noise from either known or unknown sources. A suitable approach is necessary to reconcile information from multiple sources. The Kalman filter is a widely used sequential data assimilation method for obtaining a least squares estimation of the state vector of the system [8]. It is capable to take into account various types of observations when they become available.

The Kalman filter addresses the general problem of estimating the state vector $\mathbf{S} \in \Re^n$, which represents the state of the system, including model parameters (e.g., hydraulic conductivity) and dependent variables (e.g., hydraulic head). There are two major stages in the Kalman filter procedure, forecast and assimilation [3]. The forecast stage updates the state vector $\mathbf{S}$ and the covariance matrix $\mathbf{P}$:

$$\mathbf{S}^f(i) = \mathbf{\Phi}\, \mathbf{S}^a(i-1) + \mathbf{e_1}(i), \tag{6}$$

$$\mathbf{P}^f(i) = \mathbf{\Phi}\, \mathbf{P}^a(i-1)\mathbf{\Phi}^T + \mathbf{R_1}(i), \tag{7}$$

where i is the time step, superscripts f and a stand for forecast and assimilation stages, respectively, $\mathbf{\Phi}$ is a linear transition matrix that forwards the state vector at time $i-1$ to the current time i, and $\mathbf{e_1}$ represents the process error, which is a random vector with zero mean and covariance $\mathbf{R_1}$. The forecast model will keep on running with time until new observations become available. Once new data are available, the assimilation stage starts, which can be expressed mathematically as

$$\mathbf{G}(i) = \mathbf{P}^f(i)\mathbf{H}^T \left[\mathbf{H}\, \mathbf{P}^f(i)\mathbf{H}^T + \mathbf{R_2}(i)\right]^{-1}, \tag{8}$$

$$\mathbf{d}(i) = \mathbf{H}\, \mathbf{S}^t(i) + \mathbf{e_2}(i), \tag{9}$$

$$\mathbf{S}^a(i) = \mathbf{S}^f(i) + \mathbf{G}(i) \left[\mathbf{d}(i) - \mathbf{H}\, \mathbf{S}^f(i)\right], \tag{10}$$

$$\mathbf{P}^a(i) = [\mathbf{I} - \mathbf{G}(i)\mathbf{H}]\, \mathbf{P}^f(i), \tag{11}$$

where $\mathbf{H}$ is a measurement operator that relates the state vector and observations $\mathbf{d}$, superscript T stands for matrix transpose, and $\mathbf{e_2}$ is the measurement error that has a zero mean and error covariance matrix $\mathbf{R_2}$. The first step in the assimilation stage is to compute the Kalman gain $\mathbf{G}(i)$ using (8), and then observations $\mathbf{d}$ are taken using (9). The next step is to compute a posterior estimate of the state vector in (10) by incorporating new observations, and then update a posterior error covariance from (11).

The standard Kalman filtering scheme as described above requires computing and storing the error covariance matrix of state variables, which is computationally expensive for large-scale problems (say, with millions of or even more grid nodes). To alleviate this problem, ensemble Kalman filter (EnKF), a variant of Kalman filter, has been developed [6]. In the EnKF, one keeps track of a limited number of realizations, rather than the covariance matrix as in the standard Kalman filter. At each update, the required subset of the covariance matrix is then obtained from these realizations. Owing to its conceptual simplicity, relative ease in implementation, and the ability to account for possible model noises/errors, the EnKF has been found useful in a large number of applications in various fields such as meteorology, oceanography, hydrology, and reservoir engineering [7, 27, 28, 13]. The computational cost of the EnKF is relatively low (if only a small number of realizations is used) compared to other Kalman filters. The EnKF is essentially a Monte Carlo method, depending on the number of realizations. The goodness of the covariance function approximated from an ensemble strongly depends on the number of realizations used in the ensemble, and the appropriate number of realizations may depend on the nature of the problems and may not be known *a priori*.

One major problem of various existing Kalman filtering methods is how to efficiently compute and store the covariance function in each update. Their substantial requirement on computational resources prevents us from applying them to high-resolution, large-scale simulation problems.

In this chapter, we introduce an efficient, dimension-reduced Kalman filtering scheme based on Karhunen-Loève expansions of the medium property and orthogonal polynomial decompositions of the state variable [40]. The covariance of the medium property is effectively approximated by a small number of eigenvalues and eigenfunctions using the Karhunen-Loève (KL) decomposition. The same number of first-order head fields (modes) are solved and stored in each update. Compared to the full covariance matrix in the EnKF, the finite number of head modes used to approximate it represents a significant reduction in random dimension. The reconstruction of the covariance functions from the KL decomposition of $Y(\mathbf{x})$ and the first-order head modes can be done whenever needed. In each update, the forward problem is solved using an efficient KL-based moment method [39] that gives a set of functions from which the mean and covariance of the state variables can be constructed, when needed. The statistics of both the medium property and the system responses are then efficiently updated with the available measurements at this

time using the auto- and cross-covariances obtained from the forward modeling. The KL-based Kalman filter (KLKF) will significantly reduce required computational resources (both computational time and storage) and allows us to efficiently incorporate continuous observations into high-resolution predictive models for flow and transport in large-scale problems.

4 Karhunen-Loève Decomposition

For a stochastic process $Y(\mathbf{x},\omega) = \ln K_s(\mathbf{x},\omega)$, where $\mathbf{x} \in D$ and $\omega \in \Omega$ (a probability space), the covariance function $C_Y(\mathbf{x},\mathbf{y}) = \langle Y'(\mathbf{x},\omega)Y'(\mathbf{y},\omega)\rangle$ is bounded, symmetric, and positive definite, thus it can be decomposed as [9, 10, 39].

$$C_Y(\mathbf{x},\mathbf{y}) = \sum_{m=1}^{\infty} \lambda_m f_m(\mathbf{x}) f_m(\mathbf{y}), \tag{12}$$

where m is the index of modes, λ_m and f_m are eigenvalues and deterministic eigenfunctions, respectively. The eigenvalues and eigenfunctions can be solved from the following Fredholm equation

$$\int_D C_Y(\mathbf{x},\mathbf{y}) f(\mathbf{y}) d\mathbf{y} = \lambda f(\mathbf{x}). \tag{13}$$

The set of eigenfunctions are orthogonal and form a complete set

$$\int_D f_n(\mathbf{x}) f_m(\mathbf{x}) d\mathbf{x} = \delta_{mn}, \tag{14}$$

where δ_{mn} is the Kronecker delta, $\delta_{mn} = 1$ for $m = n$ and 0 otherwise. In addition, it can be shown from (12) that the summation of all eigenvalues equals the total variability of $Y(\mathbf{x})$, $\sum_{n=1}^{\infty} \lambda_n = \int_D \sigma_Y^2(\mathbf{x}) d\mathbf{x}$. One may sort the set of eigenvalues λ_m in a non-increasing order, and the corresponding eigenfunctions then exhibit a decreasing characteristic scale as the index m increases [9, 39]. The stochastic process $Y(\mathbf{x})$ then can be expanded as:

$$Y(\mathbf{x}) = \langle Y(\mathbf{x})\rangle + \sum_{m=1}^{\infty} \xi_m \sqrt{\lambda_m} f_m(\mathbf{x}), \tag{15}$$

where ξ_m are orthogonal random variables, i.e., $\langle \xi_m \rangle = 0$ and $\langle \xi_m \xi_n \rangle = \delta_{mn}$. In the case of normally distributed Y, ξ_m are standard Gaussian random variables. It has been shown that the KL expansion, i.e., (15), is of mean square convergence and may be well approximated with a finite summation. By truncating (15), one in fact ignores some small-scale variability of Y. The number of modes required for accurately approximating $Y'(\mathbf{x})$ depends on the ratio of the correlation length to the dimension of the domain [39]. In many cases, the magnitude of eigenvalues decays very fast, which means that Y can be approximated using a small number of terms.

Note that stationarity of the process $Y(\mathbf{x})$ is not required in the above derivation. Suppose the simulation domain D is partitioned into K non-overlapping subdomains $D = \cup_{k=1}^{K} D_k$ and $D_m \cap D_n = \phi$ for $m \neq n$, where ϕ stands for the null set, and accordingly the log hydraulic conductivity field can be written as

$$Y(\mathbf{x}) = \ln K_s(\mathbf{x}) = \sum_{k=1}^{K} Y_k(\mathbf{x})\psi_k(\mathbf{x}), \tag{16}$$

where $Y_k(\mathbf{x}) = \ln K_s(\mathbf{x})$ is a spatial random function defined in subdomain D_k, and $\psi_k(\mathbf{x})$ is a deterministic indicator function given as $\psi_k(\mathbf{x}) = 1$ for $\mathbf{x} \in D_k$ and $\psi_k(\mathbf{x}) = 0$ otherwise. Lu and Zhang [25] developed an algorithm for computing eigenvalues and eigenfunctions for the nonstationary hydraulic conductivity field. The procedure can be summarized as follows:

- Equation (13) is solved for each individual zone D_k to obtain eigenvalues $\{\lambda_n^{(k)}, n = 1, 2, \cdots\}$ and eigenfunctions $\{f_n^{(k)}(\mathbf{x}), n = 1, 2, \cdots\}$;
- Extend the domain of $f_n^{(k)}(\mathbf{x})$ from D_k to the entire domain D by defining $f_n^{(k)}(\mathbf{x}) = 0$ for $\mathbf{x} \notin D_k$;
- Merge K sets of eigenvalues together $\{\lambda_n^{(k)}, k = \overline{1, K}, n = 1, 2, \cdots\}$ and sort them in a non-increasing order (denoting the sorted series as λ_k, $k = 1, 2, \cdots$);
- Arrange the set of merged eigenfunctions $\{f_n^{(k)}, k = \overline{1, K}, n = 1, 2, \cdots\}$ based on the sorted eigenvalues and denote the new set of eigenfunctions as $f_k(\mathbf{x})$, $k = 1, 2, \cdots$.

The KL decomposition of the mean-removed stochastic process $Y'(\mathbf{x})$ for both stationary and non-stationary hydraulic conductivity fields can be written as

$$Y'(\mathbf{x}) = \sum_{n=1}^{\infty} \xi_n \sqrt{\lambda_n} f_n(\mathbf{x}). \tag{17}$$

The hydraulic conductivity field conditioned on some direct measurements is a special kind of nonstationary field. Lu and Zhang [21] developed a methodology to efficiently update eigenvalues and eigenfunctions of the covariance of the hydraulic conductivity field. Since the set of eigenfunctions are complete, the basic idea of this updating scheme is to express the conditional eigenfunctions as linear combinations of the unconditional eigenfunctions and derive equations for the coefficients of these linear combinations. By doing so, the problem of finding the eigenvalues and eigenfunctions of a conditional covariance function $C_Y^{(c)}(\mathbf{x}, \mathbf{y})$ reduces to the problem of finding the eigenvalues and eigenvectors of an $N_Y \times N_Y$ symmetric matrix, where N_Y is the number of conditioning points, which in general is small. The computational cost of finding conditional eigenvalues and eigenfunctions in this way is much less than that of directly solving (13). This updating scheme is a critical component of the KLKF method described in Section 6.

5 KL-based Moment Equations

The Kalman filtering method requires to compute the covariance matrix of the system response, i.e., the $\mathbf{P}$ matrix in (6)-(11). In the EnKF, this is accomplished by running a series of forward modeling using a number of realizations. In the KLKF method, the covariance of head is derived from forward modeling using the KL-based moment method (KLME). The KLME has been described in detail in [39, 21, 22]. For completeness, a brief description is given as follows.

Since the dependent variable $h(\mathbf{x},t)$ is a function of the input variability $\sigma_Y^2(\mathbf{x})$, one may formally express $h(\mathbf{x},t)$ as an infinite series $h(\mathbf{x},t) = \sum_{m=1}^{\infty} h^{(m)}(\mathbf{x},t)$, where the order of each term in this summation is with respect to σ_Y, the standard deviation of Y. We only keep the first two terms in this series, and the KLKF method described here is based on the first-order approximation of the pressure head. By substituting the expansions of $h(\mathbf{x},t)$ and Y into (1), we obtain the governing equations for zeroth-order pressure head $h^{(0)}(\mathbf{x},t)$

$$\nabla \cdot \Big[K_G(\mathbf{x})\nabla h^{(0)}(\mathbf{x},t)\Big] + g(\mathbf{x},t) = S_s \frac{\partial h^{(0)}(\mathbf{x},t)}{\partial t}, \tag{18}$$

$$h^{(0)}(\mathbf{x},0) = H_0(\mathbf{x}), \qquad \mathbf{x} \in D, \tag{19}$$

$$h^{(0)}(\mathbf{x},t) = H(\mathbf{x},t), \qquad \mathbf{x} \in \Gamma_D, \tag{20}$$

$$K_G(\mathbf{x})\nabla h^{(0)}(\mathbf{x},t) \cdot \mathbf{n}(\mathbf{x}) = -Q(\mathbf{x},t), \qquad \mathbf{x} \in \Gamma_N, \tag{21}$$

and for the first-order pressure head term $h^{(1)}(\mathbf{x},t)$,

$$\nabla \cdot \Big[K_G(\mathbf{x})\nabla h^{(1)}(\mathbf{x},t)\Big] + g^{(1)}(\mathbf{x},t) = S_s \frac{\partial h^{(1)}(\mathbf{x},t)}{\partial t}, \tag{22}$$

$$h^{(1)}(\mathbf{x},0) = 0, \quad \mathbf{x} \in D, \tag{23}$$

$$h^{(1)}(\mathbf{x},t) = 0, \quad \mathbf{x} \in \Gamma_D, \tag{24}$$

$$K_G(\mathbf{x})\nabla h^{(1)}(\mathbf{x},t) \cdot \mathbf{n}(\mathbf{x}) = Q(\mathbf{x},t)Y'(\mathbf{x}), \quad \mathbf{x} \in \Gamma_N, \tag{25}$$

where K_G is the geometric mean of the K_s, and

$$g^{(1)}(\mathbf{x},t) = S_s Y'(\mathbf{x}) \frac{\partial h^{(0)}(\mathbf{x},t)}{\partial t} + K_G(\mathbf{x})\nabla Y'(\mathbf{x}) \cdot \nabla h^{(0)}(\mathbf{x},t) - g(\mathbf{x},t)Y'(\mathbf{x}). \tag{26}$$

Equations for higher-order mean head and head covariance can be found in [39, 21, 22, 24]

Equations (18)-(21) are the governing equations and initial and boundary conditions for the zeroth-order conditional mean head. In the conventional moment method, the equations for the first-order (in terms of σ_Y^2) head covariance can be derived from (22)-(25) upon multiplying these equations by

$h^{(1)}(\boldsymbol{\chi},\tau)$ and taking the ensemble mean. It has been shown that the conventional moment method is computationally expensive, especially for large-scale problems [22]. In fact, to solve the pressure head covariance up to the first order (in σ_Y^2) using the conventional moment-equaiton method, it is required to solve the sets of linear algebraic equations with N unknowns (N being the number of nodes in the numerical grid) for $2N$ times: N times for solving the cross-covariance C_{Yh} and N times for the covariance C_h. Solving the pressure head covariance with higher-order corrections is possible, but the computational effort is very demanding. For instance, solving the pressure head covariance up to the second order in terms of σ_Y^2 requires solving sets of linear algebraic equations with N unknowns for N^2 times.

In the KLME method, instead of dealing with $h^{(1)}$ directly from (22)-(26), we further assume that the first-order head can be exprssed as a polynomial expansion in terms of the orthogonal random variables ξ_m:

$$h^{(1)}(\mathbf{x},t) = \sum_{m=1}^{\infty} \xi_m h_m^{(1)}(\mathbf{x},t), \tag{27}$$

where $h_m^{(1)}$ are deterministic, first-order head with mode m. After substituting this expansion and the KL decomposition of the log hydraulic conductivity, i.e., (17), into (22)-(26) and recalling the orthogonality of random variables ξ_m, we obtain sets of governing equations for deterministic coefficients $h_n^{(1)}$:

$$\nabla \cdot \left[K_G(\mathbf{x})\nabla h_n^{(1)}(\mathbf{x},t)\right] + g_n^{(1)}(\mathbf{x},t) = S_s \frac{\partial h_n^{(1)}(\mathbf{x},t)}{\partial t}, \tag{28}$$

$$h_n^{(1)}(\mathbf{x},0) = 0, \quad \mathbf{x} \in D, \tag{29}$$

$$h_n^{(1)}(\mathbf{x},t) = 0, \quad \mathbf{x} \in \Gamma_D, \tag{30}$$

$$K_G(\mathbf{x})\nabla h_n^{(1)}(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x}) = Q(\mathbf{x},t)\sqrt{\lambda_n} f_n(\mathbf{x}), \quad \mathbf{x} \in \Gamma_N, \tag{31}$$

where

$$g_n^{(1)}(\mathbf{x},t) = \left[S_s \frac{\partial h^{(0)}(\mathbf{x},t)}{\partial t} - g(\mathbf{x},t)\right]\sqrt{\lambda_n} f_n(\mathbf{x}) + K_G(\mathbf{x})\sqrt{\lambda_n}\nabla f_n(\mathbf{x})\cdot\nabla h^{(0)}(\mathbf{x},t). \tag{32}$$

Here the source term $g_n^{(1)}$ and the flux bounday term in the right side of (31) represent the effect of heterogeneity of the hydraulic conductivity. Note that (28)-(32) are driving by terms that are proportional to $\sqrt{\lambda_n}$. Since $\sqrt{\lambda_n}$ decreases quickly, the first-order head $h^{(1)}$ in (27) can be approximated by a limited number of terms, say M terms. It is worthy to point out that the zeroth-order head equation and the M sets of equations for coefficients $h_n^{(1)}$ (totally there are $M+1$ sets of equations) have the exactly same structure as the original flow equation. By changing the input parameters, the KL-based moment equations can be solved easily with existing flow simulators, such as MODFLOW [15].

Once the coefficients $h_n^{(1)}$, $n = \overline{1, M}$, are solved, we only need to store an $M \times N$ matrix rather than an $N \times N$ full matrix C_h. The head covariance as well as the cross-covariance between the head and the log hydraulic conductivity, which are required in the data assimilation process, can be approximated up to first order in terms of σ_Y^2 whenever needed:

$$C_{Yh}(\mathbf{x}; \mathbf{y}, \tau) = \sum_{m=1}^{M} \sqrt{\lambda_m} f_m(\mathbf{x}) h_m^{(1)}(\mathbf{y}, \tau), \tag{33}$$

$$C_h(\mathbf{x}, t; \mathbf{y}, \tau) = \sum_{m=1}^{M} h_m^{(1)}(\mathbf{x}, t) h_m^{(1)}(\mathbf{y}, \tau). \tag{34}$$

The accuracy of the KLKF method highly depends on how well one can approximate C_{Yh} and C_h from (33)-(34). Zhang and Lu [2004] showed that, when the unconditional variability σ_Y^2 is small, these first-order approximations are close to the true solutions. However, if the variability σ_Y^2 is large (say, $\sigma_Y^2 > 2.0$), higher-order corrections are needed unless the correlation length is small. In the latter case, first-order solutions are sufficiently accurate, although in this case, the number of modes should be relatively large [*Zhang and Lu*, 2004]. In general, the number of required modes depends on the dimensionless size (in terms of the correlation length) of the domain. It is our experience that 100 modes are enough for most of cases we examined.

6 KL-based Data Assimilation Methodology

The data assimilation process consists of a series of updating steps, each of which represents the time when observations become available and/or the updating process is operated. In this section, the time symbol t is suppressed, because the discussion is based on any fixed assimilation step. Similar to the EnKF, the KLKF requires to compute covariance functions C_Y, C_{Yh}, and C_h, as presented in (12), (33), and (34). This implies that the eigenvalues and eigenfunctions, as well as the first-order term $h_n^{(1)}$ have to be updated at each assimilation step, which in turn requires updating the mean hydraulic conductivity field and the zeroth- and first-order head terms. Certainly, because of conditioning, at each updating step, the covariance function of the log hydraulic conductivity is nonstationary and its eigenvalues and eigenfunctions have to be solved numerically. It is well-known that solving eigenfunctions numerically is computationally very expensive for large-scale problems. For this reason, Zhang et al. [40] developed an algorithm that requires solving the eigenvalue problem at the first time step and then efficiently updating eigenvalues and eigenfunctions at the sequential assimilation steps.

6.1 Updating mean hydraulic conductivity field

At each update step, for the given observations described in Section 2, the updated mean and covariance of Y upon incorporating new observations can be derived from the cokriging technique:

$$Y^{(0)(c)}(\mathbf{x}) = Y^{(0)}(\mathbf{x}) + \sum_{i=1}^{N_Y} \alpha_i(\mathbf{x}) \left[Y_{obs}(\mathbf{x}_i) - Y^{(0)}(\mathbf{x}_i) \right] + \sum_{i=1}^{N_h} \beta_i(\mathbf{x}) \left[h_{obs}(\boldsymbol{\chi}_i) - h^{(0)}(\boldsymbol{\chi}_i) \right], \tag{35}$$

and

$$C_Y^{(c)}(\mathbf{x}_i, \mathbf{y}_j) = C_Y(\mathbf{x}_i, \mathbf{y}_j) - \sum_{n=1}^{N_Y} \alpha_n(\mathbf{x}_i) C_Y(\mathbf{x}_n, \mathbf{y}_j) - \sum_{n=1}^{N_h} \beta_n(\mathbf{x}_i) C_{Yh}(\mathbf{y}_j, \boldsymbol{\chi}_n), \tag{36}$$

where the quantities with (without) superscript (c) stands for the values after (before) incorporating observations at this time step, and $\alpha_i(\mathbf{x})$ and $\beta_i(\mathbf{x})$ are weighting functions, representing the relative importance of each measurement $Y_{obs}(\mathbf{x}_i)$ and $h_{obs}(\boldsymbol{\chi}_i)$ in predicting the value of $Y^{(0)(c)}(\mathbf{x})$ at location $\mathbf{x}$. The observations $Y_{obs}(\mathbf{x}_i)$ and $h_{obs}(\boldsymbol{\chi}_i)$ may include noises:

$$Y_{obs}(\mathbf{x}_i) = Y^t(\mathbf{x}_i) + \zeta_i \epsilon_Y, \quad i = \overline{1, N_Y} \tag{37}$$

$$h_{obs}(\boldsymbol{\chi}_i) = h^t(\boldsymbol{\chi}_i) + \zeta_i \epsilon_h, \quad i = \overline{1, N_h} \tag{38}$$

where $Y^t(\mathbf{x}_i)$ and $h^t(\boldsymbol{\chi}_i)$ are unknown true values at observation locations, ζ_i are Gaussian random variables with zero mean and unit variance, and ϵ_Y and ϵ_h are the standard deviations of measurements errors of the log hydraulic conductivity $Y(\mathbf{x})$ and pressure head $h(\boldsymbol{\chi})$, which are assumed to be known.

The weighting functions in (35) are solutions of the following cokriging equations:

$$\sum_{i=1}^{N_Y} \alpha_i(\mathbf{x}) C_Y(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{N_h} \beta_i(\mathbf{x}) C_{Yh}(\mathbf{x}_j, \boldsymbol{\chi}_i) = C_Y(\mathbf{x}, \mathbf{x}_j), \; j = \overline{1, N_Y}, \tag{39}$$

$$\sum_{i=1}^{N_Y} \alpha_i(\mathbf{x}) C_{Yh}(\mathbf{x}_i, \boldsymbol{\chi}_j) + \sum_{i=1}^{N_h} \beta_i(\mathbf{x}) C_h(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) = C_{Yh}(\mathbf{x}, \boldsymbol{\chi}_j), \; j = \overline{1, N_h}. \tag{40}$$

Note that coefficients α_i and β_i are location-dependent, which means that equations (39)-(40) need to be solved for N times, where N is the number of nodes in the domain. Follow [21], α_i and β_i can be directly related to the eigenvalues, eigenfunctions, and the first-order head $h^{(1)}$. Since the set of

eigenfunctions (no matter they are unconditional or conditional) is complete, $\alpha_i(\mathbf{x})$ and $\beta_i(\mathbf{x})$ can be expanded on the basis of these eigenfunctions:

$$\alpha_i(\mathbf{x}) = \sum_{k=1}^{\infty} \alpha_{ik} f_k(\mathbf{x}), \tag{41}$$

$$\beta_i(\mathbf{x}) = \sum_{k=1}^{\infty} \beta_{ik} f_k(\mathbf{x}). \tag{42}$$

Substituting (41)-(42) into (39)-(40), multiplying $f_m(\mathbf{x})$ on the both sides of the resulted equations, and integrating the equations with respect to $\mathbf{x}$ over the domain D, we obtain equations for α_{im} and β_{im}

$$\sum_{i=1}^{N_Y} \alpha_{im} C_Y(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{N_h} \beta_{im} C_{Yh}(\mathbf{x}_j, \boldsymbol{\chi}_i) = \lambda_m f_m(\mathbf{x}_j), \quad j = \overline{1, N_Y}, \tag{43}$$

$$\sum_{i=1}^{N_Y} \alpha_{im} C_{Yh}(\mathbf{x}_i, \boldsymbol{\chi}_j) + \sum_{i=1}^{N_h} \beta_{im} C_h(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) = \sqrt{\lambda_m} h_m^{(1)}(\mathbf{x}_j), \quad j = \overline{1, N_h}, \tag{44}$$

where the cross-covariance C_{Yh}, auto-covariance of pressure head C_h, and auto-covariance of the log hydraulic conductivity C_Y are given by (33), (34), and (12), respectively. All these auto- and cross-covariance functions depend on simulated time and thus need to be updated at each assimilation step. It should be noted that (43)-(44) are solved only M times (in general $M << N$), where M is the number of modes needed to approximate Y with a desired accuracy. In addition, solving α_{im} and β_{im} will facilitate updating eigenvalues and eigenfunctions, as described later.

6.2 Updating pressure head fields

The zeroth-order pressure head can be updated with both types of observations in the same manner:

$$\begin{aligned} h^{(0)(c)}(\mathbf{x}) = h^{(0)}(\mathbf{x}) &+ \sum_{i=1}^{N_Y} \mu_i(\mathbf{x}) \left[Y_{obs}(\mathbf{x}_i) - Y^{(0)}(\mathbf{x}_i) \right] \\ &+ \sum_{i=1}^{N_h} \eta_i(\mathbf{x}) \left[h_{obs}(\boldsymbol{\chi}_i) - h^{(0)}(\boldsymbol{\chi}_i) \right], \end{aligned} \tag{45}$$

where $\mu_i(\mathbf{x})$ and $\eta_i(\mathbf{x})$ are subject to:

$$\sum_{i=1}^{N_Y} \mu_i(\mathbf{x}) C_Y(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{N_h} \eta_i(\mathbf{x}) C_{Yh}(\mathbf{x}_j, \boldsymbol{\chi}_i) = C_{Yh}(\mathbf{x}_j, \mathbf{x}), \; j = \overline{1, N_Y}, \tag{46}$$

$$\sum_{i=1}^{N_Y} \mu_i(\mathbf{x}) C_{Yh}(\mathbf{x}_i, \boldsymbol{\chi}_j) + \sum_{i=1}^{N_h} \eta_i(\mathbf{x}) C_h(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) = C_h(\mathbf{x}, \boldsymbol{\chi}_j),\ j = \overline{1, N_h}. \quad (47)$$

Coefficients α_{im}, β_{im}, μ_i, and η_i in (43)-(44) and (46)-(47) can be computed through solving a set of linear algebraic equations with the same coefficient matrix. However, the condition number of this coefficient matrix could be extremely large, especially after several assimilation steps, because assimilating observations will result in a great reduction of (co)variance. Due to the ill-conditioned matrix, the truncation error will easily be amplified reflecting on the anomalies. Dietrich and Newsam [5] analyzed the cause of the ill-conditioning and proposed that adding a relaxation term (or an explicit error matrix) to the coefficient matrix can resolve this problem. In general, a relatively large relaxation term slows down the rate of convergence but with the price of losing information, while a small value may lead to numerical instability. The error matrix can be obtained through a maximum likelihood approach for the purpose of improving the conditioning and minimizing the loss of information. Yeh et al. [37] added a relaxation term to the diagonal components of the matrix to reduce the condition number of the matrix, and the relaxation term is a fraction of the maximum value of the coefficient matrix. In the example shown in Section 7, a constant relaxation term is added to the diagonal components of , which will be discussed further with illustrative examples.

Similarly to the zeroth-order head term, the first-order pressure head is updated by:

$$h^{(1)(c)}(\mathbf{x}) = h^{(1)}(\mathbf{x}) - \sum_{i=1}^{N_Y} \mu_i(\mathbf{x}) Y'(\mathbf{x}_i) - \sum_{i=1}^{N_h} \eta_i(\mathbf{x}) h^{(1)}(\boldsymbol{\chi}_i) \quad (48)$$

Because both Y' and $h^{(1)}$ can be expanded based on ξ_m, by substituting (17) and (27) into (48), instead of updating $h^{(1)}$ directly, the coefficients $h_m^{(1)}$ can be updated as:

$$h_m^{(1)(c)}(\mathbf{x}) = h_m^{(1)}(\mathbf{x}) - \sum_{i=1}^{N_Y} \mu_i(\mathbf{x}) \sqrt{\lambda_m} f_m(\mathbf{x}_i) - \sum_{i=1}^{N_h} \eta_i(\mathbf{x}) h_m^{(1)}(\boldsymbol{\chi}_i) \quad (49)$$

It can be shown that to the first order, (49) recovers the usual cokriging equation for the head covariance similar to (36).

6.3 Updating eigenvalues and eigenfunctions

Because of the non-stationality of the covariance matrix given in (36), the conditional eigenvalues and eigenfunctions have to be solved numerically. Here we follow the method in [21] with some modification to incorporate the influence of the pressure head measurements. By definition, the eigenvalues $\lambda_m^{(c)}$

and their corresponding eigenfunctions $f_m^{(c)}$ can be solved from (13) upon replacing C_Y by the conditional covariance function $C_Y^{(c)}$. Because of high computational cost in solving (13), this equation will be solved only at the first assimilation step and the eigenvalues and eigenfunctions for any sequential step will be derived from the results at the previous step. Since the set of eigenfunctions $f_m(\mathbf{x})$ computed at the previous assimilation step is complete, the eigenfunctions $f_m^{(c)}(\mathbf{x})$ at the current step (after assimilating new measurements) can be expanded in terms of $f_m(\mathbf{x})$ as:

$$f_m^{(c)}(\mathbf{x}) = \sum_{p=1}^{M} d_{mp} f_p(\mathbf{x}), \quad m = \overline{1, M} \tag{50}$$

where the coefficient matrix $\mathbf{D} = (d_{mp})_{M\times M}$ is to be determined. Substituting this expression and (36) into (13), multiplying $f_m(\mathbf{y})$ on the both sides of the derived equation, and integrating it with respect to $\mathbf{y}$ over the domain D, yields:

$$\lambda_m d_m - \sum_{k=1}^{M} \left(\sum_{i=1}^{N_Y} \alpha_{ik} \lambda_m f_m(\mathbf{x}_i) + \sum_{i=1}^{N_h} \beta_{ik} h_m^{(1)}(\boldsymbol{\chi}_i) \right) d_k = \lambda_m^{(c)} d_m, m = \overline{1, M} \tag{51}$$

It can also be expressed in a succinct matrix form as:

$$\left(\mathbf{A} - \lambda^{(\mathbf{c})} \mathbf{I} \right) \mathbf{D} = 0 \tag{52}$$

where

$$a_{km} = \lambda_m \delta_{km} - \sum_{k=1}^{M} \left(\sum_{i=1}^{N_Y} \alpha_{ik} \lambda_m f_m(\mathbf{x}_i) + \sum_{i=1}^{N_h} \beta_{ik} h_m^{(1)}(\boldsymbol{\chi}_i) \right) \tag{53}$$

are components of $\mathbf{A} = (a_{km})_{M\times M}$, $\lambda^{(c)} = diag(\lambda_1^{(c)}, \cdots, \lambda_M^{(c)})$, and $\mathbf{I}$ is an $M \times M$ identical matrix. Therefore, the problem of finding the eigenvalues and eigenfunctions of a nonstationary covariance matrix $C_Y^{(c)}(\mathbf{x}, \mathbf{y})$ of size $N \times N$ reduces to the problem of finding the eigenvalues $\lambda^{(c)}$ and eigenvectors d of an $M \times M$ matrix for M times, where N is the number of grid nodes, and M is the number of modes. Note that the number of grid nodes N is usually much larger than the number of modes M. Once $\mathbf{D}$ is solved, conditional eigenvector $f_m^{(c)}$ corresponding to each conditional eigenvalue $\lambda_m^{(c)}$ can be constructed using (50). The updated eigenvalues and eigenvectors as well as the updated pressure head terms in (45) and (49) are fed into the KLME forward model. Since directly solving the Fredholm equation is computationally expensive, the proposed algorithm for updating conditional eigenvalues and eigenfunctions at each updating step has a significant advantage, and we only need to solve the Fredholm equation at the first time step rather than at each assimilation step.

6.4 Karhunen-Loève based Kalman filter

The methodology described in the previous section has a different format than the traditional Kalman filter, where the Kalman gain and the updating step are formulated as:

$$\mathbf{G} = \mathbf{P}^f\mathbf{H}^T[\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}]^{-1} \tag{54}$$

$$\mathbf{S}^a = \mathbf{S}^f + \mathbf{G}[\mathbf{d} - \mathbf{H}\mathbf{S}^f] \tag{55}$$

where $\mathbf{K}$ is the Kalman gain, $\mathbf{H}$ is the observation operator, $\mathbf{R}$ is the observation error covariance matrix, $\mathbf{d}$ is the observation vector, $\mathbf{S}^f$ is the forecast state vector, $\mathbf{S}^a$ is the updated state vector, and $\mathbf{P}^f$ is the covariance matrix of the forecast state vector. In fact, the cokriging based updating scheme is exactly the same as the traditional Kalman filter. If we define the state vector as $\mathbf{S} = (Y(\mathbf{y}_1), \cdots, Y(\mathbf{y}_N), h(\mathbf{y}_1), \cdots, h(\mathbf{y}_N))^T$, the Kalman gain can be constructed with the coefficients $\alpha_i(\mathbf{x})$, $\beta_i(\mathbf{x})$, $\mu_i(\mathbf{x})$, and $\eta_i(\mathbf{x})$ described before:

$$\mathbf{G} = \begin{bmatrix} G_\alpha & G_\beta \\ G_\mu & G_\eta \end{bmatrix}_{2\times N, N_Y+N_h} \tag{56}$$

where $\mathbf{G}_\alpha = (\alpha_1, \cdots, \alpha_{N_Y})$, $\mathbf{G}_\beta = (\beta_1, \cdots, \beta_{N_h})$, $\mathbf{G}_\mu = (\mu_1(, \cdots, \mu_{N_Y})$, and $\mathbf{G}_\eta = (\eta_1, \cdots, \eta_{N_h})$. Note that each of α_i, β_i, μ_i, and η_i is a vector of $N \times 1$.

The major difference between the KLKF approach and the cokriging method is that the KLKF is a sequential or inline method, while cokriging is a statistical interpolation method, which only operates at a fixed time. Sometimes cokriging is performed iteratively to account for the possible nonlinear effects in order to obtain a reasonable estimation, but the iteration is still based on a certain time [37]. The KLKF uses the KLME method for advancing the system with time, and incorporates the observations at the time when they become available and update the system at the same time. After the current assimilation step, the updated system responses are taken as the initial conditions for the next forward step. The updated model is then run until the next set of observations become available, at which the updating step will be performed again.

7 Illustrative Examples

In this section, a synthetic two-dimensional example is used to demonstrate the applicability of the KLKF method in estimating the hydraulic conductivity field by assimilating both pressure head and hydraulic conductivity measurements. The results are then compared with those from the EnKF method in terms of both the computational cost and accuracy [40].

The flow domain is a square of size $L_x = L_y = 800$[L](where L is any consistent length unit), uniformly discretized into 40×40 square elements, as shown in Figure 1. A pumping well and an injection well are placed at (240[L], 160[L]) and (540[L], 560[L]), respectively, with a volumetric flow rate

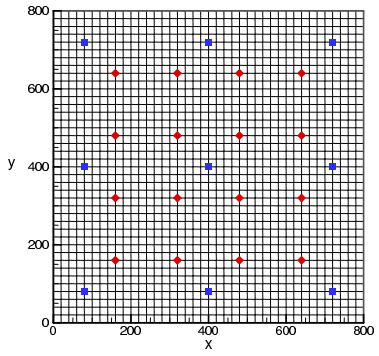

Fig. 1. The flow domain and the observation locations for $\ln K_s$(9 blue squares) and h (both blue and red symbols).

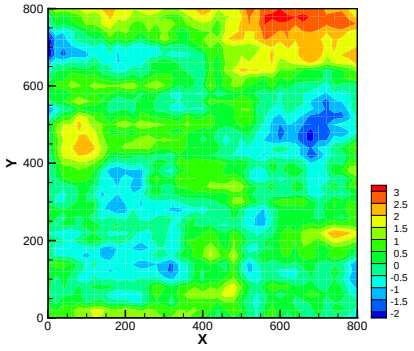

Fig. 2. The reference log hydraulic conductivity field

of 150 [L^3/day]. Both wells are active throughout the entire simulation period with constant flow rates. The two lateral boundaries are no-flow boundaries, while the left and the right are Dirichlet boundaries with prescribed pressure head of 202 [L] and 198 [L], respectively. Storage coefficient is assumed to be a constant and taken as 0.0001. The log hydraulic conductivity field is treated as a spatially-correlated Gaussian random field with zero mean and unit variance. The unconditional hydraulic conductivity field is also assumed to be second-order stationary characterized by a separable exponential covariance function

$$C_Y(\mathbf{x}_1, \mathbf{x}_2) = C_Y(x_1, y_1; x_2, y_2) = \sigma_Y^2 \exp\left[\frac{|x_1 - x_2|}{\lambda_x} - \frac{|y_1 - y_2|}{\lambda_y}\right], \qquad (57)$$

where $\sigma_Y^2 = 1.0$ is the unconditional variance of Y, and $\lambda_x = 200$ [L] and $\lambda_y = 100$ [L] are correlation lengths in x and y directions, respectively.

In this synthetic example, an unconditional realization of the log hydraulic conductivity field is generated under given statistics using a random field generator based on the KL decomposition [39]. This field is then considered as the "true" field, called the reference field, as shown in Figure 2. Nine samples are taken from this reference field at selected locations as shown in Figure 1 (in blue) and these samples are considered as direct measurements of the log hydraulic conductivity field. A forward transient simulation is conducted using the reference hydraulic conductivity field. For this model setup, the fluid flow reaches steady state at about $t = 10$ [day]. This period is chosen as the duration of the total simulation time, which is then subdivided into 50 equally-sized time intervals with a size of 0.2 [day]. Twenty-five pressure head measurements are then taken at selected locations (both blue and red points in Fig. 1) at elapsed time $t = 0.2 + 0.6k$, $k = 0, 1, \cdots, 16$. It is assumed

that the hydraulic conductivity measurements are error-free, while the pressure head observations are noisy and the measurement error follows a normal distribution $N(0, 2.5 \times 10^{-3})$. The hydraulic conductivity measurements are assimilated with the first set of pressure head measurements at $t = 0.2$ [day], and after that only pressure head measurements are assimilated every 0.6 day.

The statistics (mean, variance, and correlation lengths) of the unknown hydraulic conductivity field are usually inferred from its direct measurements. It is often that the number of measurements is not sufficient to infer the true statistics. As a consequence, in the KLKF method, the initial statistics of the hydraulic conductivity field are taken as $\langle Y \rangle = 0.0$, $\sigma_Y^2 = 1.2$, $\lambda_x = 220$ [L], and $\lambda_y = 120$ [L], which are slightly different from the those statistics used in generating the reference field. Initial eigenvalues and eigenfunctions are solved using these statistics, and the impact of the number of modes has been explored by approximating Y and $h^{(1)}$ with 50, 100, and 200 modes. For the purpose of comparison, the EnKF method is also implemented with the same initial statistics as in the KLKF method, and results from the EnKF with 100, 200, and 1000 realizations are compared with those from the KLKF to assess the accuracy and efficiency of the KLKF method.

7.1 Comparison of Accuracy

Two measures are commonly used to evaluate the accuracy of the Kalman filtering schemes. The root mean square error (RMSE) is used to compare the estimated Y field with the reference field, which is given as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[Y^*(\mathbf{x}_i) - Y^t(\mathbf{x}_i)\right]^2}, \tag{58}$$

where Y^* stands for the estimated value, Y^t stands for the true reference value, and N is the number of grid nodes.

Another measure is ensemble spread, which is the averaged variation of the ensemble and is defined as

$$Spread = \sqrt{\frac{1}{N}\sum_{i=1}^{N} VAR(\mathbf{x}_i)} \tag{59}$$

where $VAR(\mathbf{x}_i)$ is the ensemble variance. If the EnKF estimates the uncertainty of the state vector properly, the ensemble spread should be close to the RMSE.

Figure 3(a) compares the RMSE of the KLKF and the EnKF methods with a various number of realizations (or modes). The KLKF is operated with a relaxation term $\epsilon = 0.3$ for the cases with 50, 100, and 200 modes. It is seen from the figure that with 100 principal modes the KLKF can properly propagate the statistics of the real randomness, and the estimate is as good

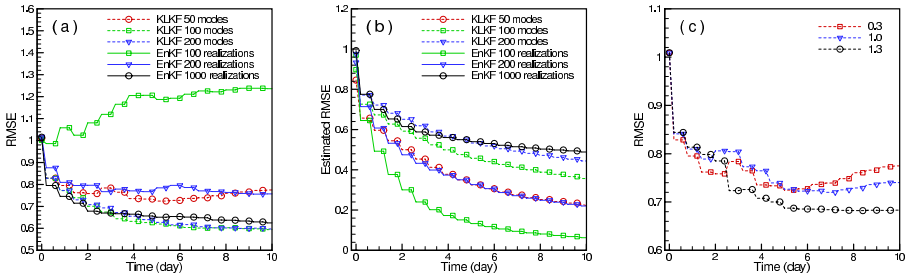

Fig. 3. (a) RMSE and (b) Ensemble spread of the KLKF and the EnKF as a function of time for various number of modes and ensemble sizes.

as that from the EnKF method with 1000 realizations. The result from only 50 modes is not very stable, but it is comparable with the EnKF with 200 realizations. The KLKF with 100 and 200 modes give similar behavior of the RMSE, which means that adding more modes beyond 100 modes may not significantly improve the results. It is important to note that for this case the EnKF with 100 realizations begins to diverge after the first assimilation step.

Figure 3(b) illustrates the ensemble spread of the EnKF with respect to time, in comparison to the uncertainty estimated by the KLKF. Comparing the values shown in Figure 3(b) with the corresponding values in Figure 3(a), the ensemble spread systematically underestimates the real deviation due to the limited size of the ensemble numbers. For instance, the EnKF with 100 realizations fails to reduce the RMSE after the first assimilation step, while the ensemble spread keeps decreasing, indicating that the ensemble is converging to a wrong solution. The estimated uncertainty of the KLKF with 200 modes is equivalent to that of the EnKF with 1000 realizations, which results in the similarity of their corresponding RMSE shown in Figure 3(a). However, the RMSEs of the KLKF with 200 modes and 100 modes are close, and the KLKF with 200 modes has a better estimate of the real deviation, which shows the potential to further incorporate new observations.

The relaxation term used in the KLKF reduces the condition number of the coefficient matrix in (43)-(44) and (46)-(47), and therefore improves stability of the KLKF solution. Numerical experiments [40] show that the performance of the KLKF is sensitive to the choice of relaxation term when the number of modes is small, for instance 50. However, reasonable results can be achieved by cautiously choosing the relaxation term for the KLKF with 50 modes (shown in Figure 3(c)). With larger number of modes, the KLKF has satisfactory performance as long as the relaxation term is changing within a reasonable range. Sensitivity of model results on the relaxation term has been investigated in [40]. In the KLKF approach, the log hydraulic conductivity measurements are used only at the first assimilation step, and the relaxation term is only added to the diagonal terms of the pressure head covariance matrix. If the

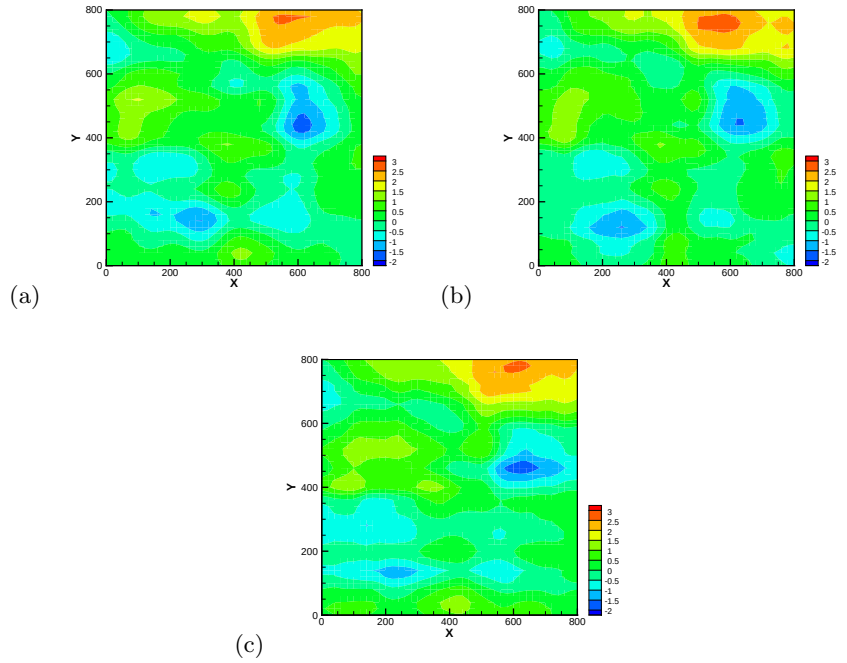

Fig. 4. The estimated mean $\ln K_s$ fields (a) KLKF with 100 modes, (b) KLKF with 200 modes, and (c) EnKF with 1000 realizations.

direct measurements are assimilated with the pressure head observations at every assimilation step, it is expected that a relaxation term needs to be added to the diagonal components of both log hydraulic conductivity and pressure head covariance matrices.

Zhang et al. [40] also investigated the impact of incorrect initial statistics (comparing to those used in generating the reference field) on the final assimilation results. Their study indicates that the impact is minor if the number of modes included in the KLKF method is enough, for example, 100 modes.

The contours of the log hydraulic conductivity fields estimated from the KLKF method with different numbers of modes are shown in Figure 4. Also compared in the figure is the estimated field using the EnKF method with 1000 realizations. Compared to the reference field (Fig. 2), the contours from the KLKF are smoother. The KLKF with 50 modes (not shown here) fails to catch the right locations of the major pattern of the reference field, while the KLKF with 100 modes almost identifies every primary structure. The contour of the EnKF also recovers the major pattern with more details than the KLKF fields. However, these small features do not exactly replicate those of the reference field. The contours of the associated log hydraulic conductivity variance are shown in Figure 5. Comparing the KLKF results with 100 modes

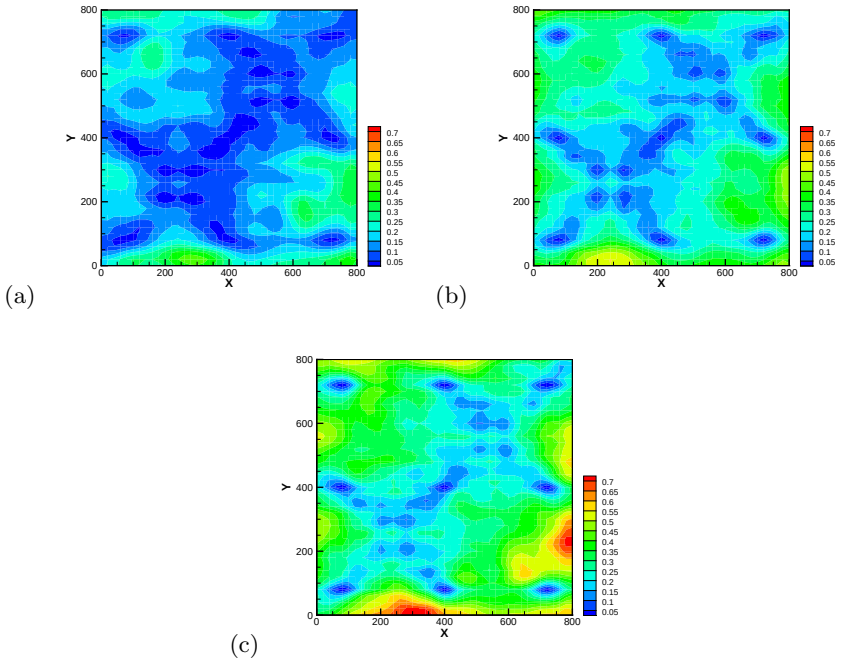

Fig. 5. The estimated $\ln K_s$ variance (a) KLKF with 100 modes, (b) KLKF with 200 modes, and (c) EnKF with 1000 realizations.

and 200 modes reveals that the variance of the estimated $\ln K_s$ field is higher if more modes are included in data assimilation. The major patterns of estimated fields from the KLKF and the EnKF are close. For the EnKF, although the log hydraulic conductivity measurements are only assimilated at the first assimilation step, the variances remain lowest at those locations. However, this feature is not very obvious in the KLKF, which may be attributed to the effect of the relaxation term.

Figure 6 compares the reference field and the estimated log hydraulic conductivity field from the KLKF with 100 modes at several elapsed times. The figure shows that with time (and available observations), the estimated field becomes closer to the reference field.

In order to test the predictability of the model using the estimated hydraulic conductivity field, we run a deterministic flow simulation from the t = 0 [day] to t = 20 [day] using the final estimated log hydraulic conductivity field (at 10th day) from the KLKF with 100 modes as the initial input. The calibrated pressure head at t = 5 [day] and the predicted head at t = 20 [day] are illustrated in Figure 7 (dashed contour lines) as compared to the reference pressure head field (solid contour lines), which is computed using the reference hydraulic conductivity field. Apparently, both head fields match very well.

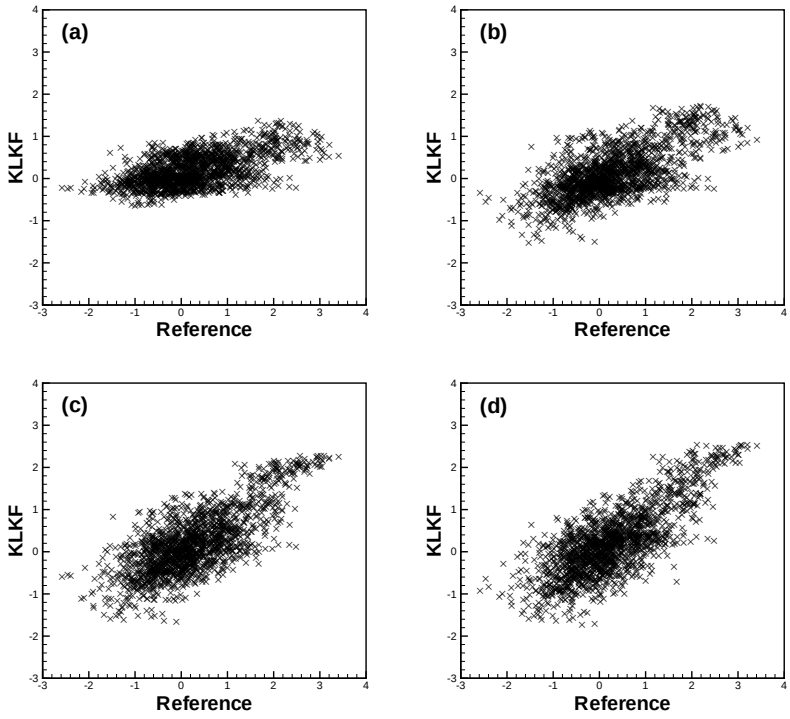

Fig. 6. Comparison between the estimated $\ln K_s$ from the KLKF with 100 modes and the reference at different times. (a) 0.2 day, (b) 2.0 day, (c) 5.0 day and (d) 10.0 day.

7.2 Computational Efficiency

If M modes are used in the KLKF, the flow equation needs to be solved for $M+1$ times (once for the zeroth-order head term $h^{(0)}(\mathbf{x})$ and M times for the first-order head terms $h_m^{(1)}(\mathbf{x})$), while for the EnKF with the ensemble size of K, the similar equation needs to be solved for K times. Since the CPU time for solving the pressure head term in the KLKF is about the same as that for one realization in the EnKF method, the computational efficiency for the two methods simply depends on how many modes in the KLKF method (or realizations in the EnKF method) are needed to approximate the statistics of the state. The illustrative example showed that the KLKF can achieve satisfactory estimation with a relatively small computational cost.

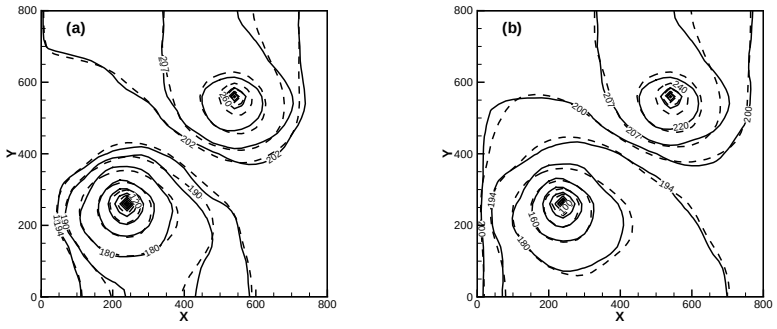

Fig. 7. Comparison of the pressure head fields computed using the reference $\ln K_s$ field (solid lines) and those computed using the estimated mean $\ln K_s$ field from K LKF with 100 modes (dashed lines). (a) pressure head at day 5 and (b) pressure head at day 20.

8 Summary and Conclusion

The Kalman filter based sequential data assimilation methods have been widely used in solving the inverse problem recently. These methods are capable of updating the system parameters continuously and sequentially with the availability of the measurements of the system responses. In these methods, both the best estimate and the corresponding uncertainty are advanced with time. A major problem associated with these existing Kalman filter based methods is the high computational cost in updating the state error covariance matrix. In this chapter, the Karhunen-Loève based Kalman filter (KLKF) is introduced. The hydraulic conductivity field is treated as a random spatial function and is decomposed using the KL expansion. The pressure head is expanded using the perturbative polynomial expansion. On the basis of these expansions, the higher-order terms are truncated and the KLKF is based on the first-order approximation of the pressure head. The KLKF utilizes only a small number of principal modes to propagate the statistics of the state vector, which greatly reduces the computational cost. The forward step can be solved accurately and efficiently using the Karhunen-Loève based moment-equation method (KLME), which can be solved in parallel using any existing flow model. The data assimilation step is operated based on the state statistics given by the forward step and the observations. A synthetic two-dimensional example shows that the KLKF method is better than the EnKF method in terms of both computational efficiency and accuracy. The example indicates that the estimated conductivity field using the KLKF method with 100 modes is reasonably close to the true reference conductivity field and the pressure head predicted using the estimated field agrees well with the reference head field.

9 Nomenclature

C_h = covariance of pressure head h
C_Y = covariance of $Y = \ln K_s$
C_{Yh} = cross-covariance between Y and h
D = the simulation domain (or its size)
D_m = subdomain
$\mathbf{d}$ = observation vector
$\mathbf{e_1}, \mathbf{e_2}$ = error vectors
$f, f_m, f_n^{(k)}$ = eigenfunction
$\mathbf{G}$ = Kalman gain
g = source/sink
$g_n^{(k)}$ = n^{th} mode of the k^{th}-order source/sink
H = prescribed head at constant head boundary
H_0 = initial head in the domain
$\mathbf{H}$ = observation operator
$h^{(k)}$ = k^{th}-order pressure head
$h_n^{(k)}$ = n^{th} mode of the k^{th}-order pressure head
$\mathbf{I}$ = identical matrix
K = number of subdomains
K_s = hydraulic conductivity
K_G = geometric mean of the hydraulic conductivity
M = number of modes
N = number of grid nodes
N_Y = number of log hydraulic conductivity observations
N_h = number of pressure head observations
$\mathbf{n}$ = an outward unit vector normal to external boundaries
$\mathbf{P}$ = covariance matrix of the state vector
$\mathbf{R_1}, \mathbf{R_2}$ = covariance matrix of error vectors $\mathbf{e_1}$, $\mathbf{e_2}$
$\mathbf{S}$ = state vector
S_s = specific storage
t = time
$\mathbf{x}, \mathbf{y}$ = Cartesian coordinate vectors
$\mathbf{x}_i$ = measurement locations of hydraulic conductivity
Y = log hydraulic conductivity
Y' = zero-mean fluctuation of Y
Y_k = log hydraulic conductivity in zone k
Γ_D = Dirichlet boundary segments
Γ_N = Neumann boundary segments
$\mathbf{\Phi}$ = linear transfer matrix
α_i, β_i = weight coefficients in cokriging estimate of Y
δ_{mn} = Kronecker data
ϵ_Y, ϵ_h = standard deviation of measurments errors on Y and h
ζ = standard Gaussian random variable
$\lambda, \lambda_m, \lambda_n^{(k)}$ = eigenvalue
λ_x, λ_y = correlation lengths in x and y directions
μ_i, η_i = weight coefficients in cokriging estimate of h
ξ_m = orthogonal Gaussian random variables
σ_Y^2 = variance of the log hydraulic conductivity
$\boldsymbol{\chi}_i$ = head measurement locations
ψ_k = indicator function for zone k

Subscript
Obs = observation
en = ensemble

Superscript
(c) = conditioned
f = forecast
u = updated
T = transpose
t = true

References

1. Anderson J L, Anderson S L (1999) Monthly Weather Review 127: 2741-2758
2. Carrera J, Neuman S P (1986) Water Resour. Res. 22(2): 199-210
3. Chen Y, Zhang D (2006) Advances in Water Resour. 29: 1107-1122
4. Dagan G (1985) Water Resour. Res. 21(1): 65-72
5. Dietrich C R, Newsam G N (1989) Stoch. Hydrol. Hydraul. 3: 293-316
6. Evensen G (1994) J. Geophys. Res. 99(C5): 10143–10162
7. Evensen G (2003) Ocean Dynamics 53: 343-367
8. Gelb A (1974) Applied Optimal Estimation, The MIT press
9. Ghanem R, Spanos P D (1991) Stochastic Finite Elements:A Spectral Approach, Springer-Verlag, New York
10. Ghanem R (1998) Water Resour. Res. 34:2123-2136
11. Ginn T R, Cushman J H (1990) Stoch. Hydrol. Hydraul. 4: 1-26
12. Gomez-Hernanez J J, Sahuquillo A S, Capilla J E (1997) J. Hydrol. 203: 162-174
13. Gu Y, Oliver D S (2004) SPE 89942
14. Gutjahr A, Wilson J (1989) Transport in Porous Media 4: 585-598
15. Harbaugh A W, Banta E R, Hill M C, and McDonald M G (2000) MODFLOW-2000, the U.S. Geological Survey modular ground-water model – User guide to modularization concepts and the Ground-Water Flow Process: U.S. Geological Survey Open-File Report 00-92.
16. Heemink A W, Verlaan M, Segers A J (2001) Monthly Weather Review 129: 1718-1728
17. Houtekamer P L, Mitchell H L (1998) Monthly Weather Review 126: 796-811
18. Hughson D L, Gutjahr A (1998) Stoch. Hydrol. and Hydraul. 12: 155-170
19. Kitanidis P K, Vomvoris E G (1983) Water Resour. Res. 19(3): 677-690
20. Liu N, Oliver D S (2005) J. Petrol. Sci. and Eng. 47: 147-161
21. Lu Z, Zhang D (2004) Adv. in Water Resour. 27(9): 859-874
22. Lu Z, Zhang D (2004) SIAM J. Sci. Comp. 26(2):558-577
23. Lu Z, Higdon D, Zhang D (2004) Comput. Methods in Water Resour., 1273-1283.
24. Lu Z, Zhang D (2006) SPE Journal 11(2), 10.2118/93452-PA, 239-247
25. Lu Z, Zhang D (2007) Multiscale Modeling and Simulation, in press
26. McLaughlin D, Townley L R (1996) Water Resour. Res. 32(5): 1131-1162
27. Mclaughlin D (2002) Adv. in Water Resour. 25: 1275-1286
28. Naevdal G, Johnsen L M, Aanonsen S I, Vefring E H (2003) SPE 84372
29. Neuman S P (1980) Water Resour. Res. 16(2): 331-346

30. Pham D, Verron J, Roubaud (1998) J. Mar. Syst. (16):323-340
31. Sun N-Z, Yeh W W-G (1992) Water Resour. Res. 28(12): 3269-3280
32. Tartakovsky D M, Neuman S P, Lu Z (1999) Water Resour. Res. 35(3): 731-745
33. Verlaan M, Heemink A W (1997) Stoch. Hydrol. Hydraul. 11:347-368
34. Whitaker J S, Hamill T M (2002) Monthly Weather Review 130: 1913-1924
35. Woodbury A D, Ulrych T J (2000) Water Resour. Res. 36(8):2081-2093
36. Yeh W W G (1986) Water Resour. Res. 22(2): 95-108
37. Yeh T C, Jin M, Hanna S (1996) Water Resour. Res. 32(1): 85-92
38. Zhang D (2002) Stochastic Methods for Flow in Porous Media: Coping with uncertainties, Academic Press, San Diego, Calif.
39. Zhang D, Lu Z (2003) J. Comput. Physics 194(2): 773–794
40. Zhang D, Lu Z, Chen Y (2007) SPE Journal, 12(1), 10.2118/95277-PA, 108-117.
41. Zimmerman D A, de Marsily G, Gotway C A, Marietta M G, Axness C L, Beauheim R L, Bras R L, Carrera J, Dagan G, Davies P B, Gallegos D P, Galli A, Gomez-Hernandez J, Grindgrod P, Gutjahr A L, Kitanidis P K, LaVenue A M, McLaughlin D, Neuman S P, RamaRao B S, Ravenne C, Rubin Y, (1998) Water Resour. Res. 34(6): 1373-1413